DOCUMENT RESUME

ED 318 803 TM 014 937

AUTHOR Linacre, John M.
TITLE Modelling Rating Scales.
PUB DATE Apr 90
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Factor Analysis; *Item Response Theory; Likert Scales; Mathematical Models; *Rating Scales; Test Construction
IDENTIFIERS Rasch Model

ABSTRACT

Determination of the intentions of the test developer is fundamental to the choice of the analytical model for a rating scale. For confirmatory analysis, the developer's intentions inform the choice of the general form of the model, representing the manner in which the respondent interacts with the scale; these intentions also inform the choice of the precise statement of that form, representing the intention of the analyst to construct, for example, an "equal-interval" scale. The nature of the Likert rating scale is discussed. Examples of general forms and precise statements are given. Forms of the measurement model discussed include the dichotomous case, the Andrich model for holistic scales, the Glas model for incremental scales, and the McCullagh model for incidental scales. Means of modeling and communicating the intentional form of the scale are outlined. Other issues addressed include incorporating the developer's intentions, anchoring the scale calibrations, and general structural concepts. A sample design problem is presented. Two tables and 16 figures are included. (TJH)

Modelling Rating Scales

by

John M. Linacre

MESA Psychometric Laboratory
Department of Education
University of Chicago

Paper presented at
American Educational Research Association Annual Meeting
Boston, Massachusetts
April 1990.

Abstract

Determination of the intention of the test developer is fundamental to the choice of the analytical model for a rating scale. For confirmatory analysis, they inform the choice of the general form of the model, representing the manner in which the respondent interacts with the scale, and also of the precise statement of that form, representing the intention of the analyst to construct, say, an "equal-interval" scale. Examples of general forms, and precise statements are given.

Key words: Rating Scale; Rasch Measurement

Introduction: The nature of a Likert rating scale

The construction of a rating scale is rarely haphazard, but is rather the result of careful thought by the test developer, who has in mind a firm idea of the manner in which the categories represent different levels of the intended variable of knowledge, attitude, experience etc. For instance, the test developer may intend to construct an equal-interval scale in order to study, say, "completion of homework assignments". Then Figure 1 portrays the test developer's intention. The categories of the rating scale are equal spaced reflecting the intention that the distances between the categories represent equal distances in the function of the category definitions. As Likert (1932) demonstrates, this straight-forward conceptualization has considerable merit.

Closer examination of the scale, however, reveals that this approach may be somewhat naive. Some children who are rated "None" may regularly almost complete an assignment or two; other children, also rated "none", may never have started even one assignment. We are forced to the conclusion that each extreme category of the scale must represent an infinite continuum of performance above or below the scale, and further that it is unlikely that our definition of the intermediate categories is exactly equal-interval.

Figure 2 portrays graphically what this might mean in terms of actual zones of performance on the necessarily infinite continuum of the underlying variable. Each extreme category (1 or 5 in this case) represents a conceptually infinite range of performance. Intermediate categories (2, 3 or 4 in this case) represent ranges of performance for which the end-points are determined by the definitions of the adjacent categories. Empirically, no two intermediate categories will have ranges of exactly the same length on the infinite variable.

Consideration of rating scale observations has often proceeded in one of two directions. Much analysis ignores the nature of the infinite continuum and follows Likert in treating the ratings themselves as an equal-interval scale. This frequently produces useful, if approximate results, but can be misleading if the observations are not towards the center of the scale, or the scale layout is far from uniform. Other analysis has appreciated the infinite nature of the continuum, but has proceeded in an exploratory manner, allowing the idiosyncracies of the data to mandate the form of the scale.

This paper is a step towards resolving this conflict between the developer's

intention and the empirical realization of the rating scale, suggesting that a confirmatory approach to rating scale be considered, one in which the developer's intentions play the major role in the analysis.

Following this line of reasoning, the problem of modelling rating scales thus has several aspects:

1) the measurement problem:
What is the general form of the measurement model which gives objective calibrations based on the developer's conception of the data?

2) the problem of intention:
In constructing the scale, the developer had some intention, usually expressed in terms such as "the categories should be equally spaced". In what way can these intentions be expressed as a special form of the measurement model, so that calibrations based on the developer's intentions can be obtained, together with fit statistics and other diagnostic information as to the extent to which the empirical data reflects those intentions ?

3) the communication problem:
How can interaction between the developer, the user and the data be promoted so that the greatest advantage is taken of both the information provided by the developer and that provided by each particular, but somewhat idiosyncratic, set of observations.

I. The form of the measurement model

1. The dichotomous case

The simplest form of the rating scale is the dichotomous item, for which the scale has merely two categories, 0 and 1. For convenience of conceptualization, we can think of a person responding to a test item with 1 (a "success") representing a higher performance level than 0 (a "failure").

The production of such ordinal observations (counts) into linear from a linear combination of the underlying parameters requires a model of the form:

$$F(P_{ni}) = B_n - D_i \qquad (1)$$

where

P_{ni} is the probability of success of person n on item i
B_n is a parameter representing the ability measure of person n, where n=1,N
D_i is a parameter representing the difficulty calibration of item i, where i=1,L.
F() is a function which monotonically transforms a value in the range (0,1) into a value in the range -infinity to +infinity.

The precise form of function F() for the dichotomous case, of which a general shape is pictured in Figure 3, has implications for the principles underlying the modelling of rating scales.

The measurement model derived by Rasch can be expressed as

$$\log(P_{ni}/(1-P_{ni})) = B_n - D_i \qquad (2)$$

in which the estimation is based on the ratio of the probability of success and the probability of failure, the logarithm of the odds.

An alternative model which would appear to be just as useful is:

$$\tan^{-1}(((P_{ni} - (1-P_{ni}))\pi/2) = B_n - D_i \qquad (3)$$

in which the comparison is based on the difference between the probability of success and the probability of failure.

However, though both models express each component (person or item) by one parameter, and are of linear form, there is a fundamental difference in their statistical properties. In Rasch's model, the person parameters can be conditioned out of the estimation of the item parameters, and vice versa. Thus, ignoring the statistical bias introduced by the possibility of extreme scores, it is not necessary to know which particular persons answered an item correctly in order to estimate the item difficulty. The margins of the response matrix, the raw scores, are sufficient statistics for estimating the parameters.

For the inverse tangent model, no sufficient statistics exist, so that, in order to estimate the parameters, it is necessary to know the details of the responses made. This threatens the basic concept of useful measurement, that the measure be essentially independent of the details of the device used to obtain it.

For a polytomous rating scale, this concept becomes yet more complex, because there appear to be many alternative, yet reasonable, ways to express a rating scale. Three are considered here, all of which can be developed from the Rasch model for the dichotomous case, but with different hypotheses about the nature of the measurement situation.

2. The Andrich Model for wholistic scales

Following Andrich (1978), the model built most closely on the work of Rasch can be expressed as

$$\log\ (P_{nij}/P_{nij-1}) = B_n - D_i - F_j \qquad \text{for } j=1,J \qquad (4)$$

where

P_{nij} is the probability of an observation in category j
P_{nij-1} is the probability of an observation in category j-1
B_n is the ability of person n
D_i is the difficulty of item i
F_j is the step difficulty or threshold between categories j-1, where the categories are numbered, say, 0,J, where, for the purposes of this discussion, all items have the same category structure.

Conceptually, this model requires that the relationship between any two adjacent categories is a dichotomous Rasch model. For J=1, the Andrich model becomes the

Rasch dichotomous model. The conceptual underpinning of the model is that each category represents a qualitatively different level of the variable, but that comprehension of all levels is required in order to place the person in any one of them. Merging adjacent categories in the data together into one category, or splitting a category into two adjacent categories, necessarily changes the meaning of all the categories, and so the frame of reference of all the parameters. For estimates from data which are in accord with the Andrich model, collapsing categories lessens the discrimination of the measuring system and so contracts the estimates towards the mean, establishing a new frame of reference. Thus, even though the category thresholds are parameterized independently, they must be considered together when interpreting them.

The sufficient statistics for the person measures and item calibrations are the margins of the score matrix, and the sufficient statistics for the category parameters are the gross counts of the number of responses observed in each category. Details of particular ratings are not need for estimation, though they are required for analysis of the fit of the data to the model, as always. A further important feature of this model is that, if the categories are reversed, i.e. counted from the other end of the scale, the measures are merely reversed in sign. Since the direction of a scale is arbitrary, this is an essential feature for measurement.

The unavoidable redefinition of the frame of reference when the rating scale is amended motivates an exploration for alternative models which allow for the addition or removal of categories without grossly disturbing the parameter estimates.

3. The Glas model for incremental scales

Glas and Verhelst (1989) present a "steps" model for rating scales also based on the Rasch dichotomous model. The rating scale item is conceptualized as a multi-stage testing item, in which success on the previous category is required before a person is considered to have attempted the next higher category. This model can be written as:

$$\log(P_{nij}/(1-P_{nij})) = B_n - D_i - F_j \qquad \text{for } j=1,J \text{ when } X_{ni}>=j-1 \qquad (5)$$

where

X_{ni} is the observation resulting from person n interacting with item i.

Each item is thus considered to be a sequence of notional category-items. The easiest category-item is administered first, followed by successively more difficult category-items are administered until either the person fails a category-item or the sequence is exhausted. Table 1 depicts the ways in which the possible responses on a scale consisting of the 4 categories 0,1,2 and 3 are decomposed into category-items.

For J=1, Glas's model is also the Rasch dichotomous model. But since each category-item is modelled to fit the Rasch dichotomous model, local independence is required to exist, conceptually, across the category-items comprising each

rating scale item. Consequently estimates of measures for the Glas model can be obtained using any software for estimation of the Rasch dichotomous model which allows missing data. If dependency between category-items exists because of the sequencing, then this will be reflected in the fit statistics. Though sufficient statistics exist for this model, the form of the data is such that fully conditional estimation fails.

The decomposition of the rating scale into category-items, expressed in Table 1, is strongly directional and not reversible without changing the meaning of the frame of reference and the calibrations in a comprehensive manner. The higher up the rating scale a person scores, the more category-items were encountered and so the more information is obtained. Reversing the category numbering would result in the person being analyzed on a test of different length. For scales in which the direction of numbering of the categories is arbitrary, Glas's model would give ambiguous results.

For scales, however, which are not wholistic, but rather incremental, Glas's model offers the possibility of splitting or merging the top category without changing the meaning of the scale. It is not necessary to know anything of the higher categories in order to interpret the meaning of the lower ones.

3) The McCullagh model for incidental scales

McCullagh (1980) presents the "proportional odds" model for rating scales in a number of versions. The version which is of interest here is that which is analogous to the Rasch dichotomous model. This model can be written as:

$$\log\left(\sum_{k=j}^{J} P_{nik} \Big/ \sum_{k=0}^{j-1} P_{nik}\right) = B_n - D_i - F_j \qquad \text{for } j=1,J \qquad (6)$$

Thus every category boundary is considered to be equivalent to a dichotomous item, not just for the adjacent categories, as in the Andrich model, but for all the categories. The rating scale is conceptualized as being based on parallel logistic ogives, Figure 4, rather than the non-parallel ogival shapes resulting from Andrich's model (cf. Figure 7).

For J=1, this is also the Rasch dichotomous model. For polytomous scales, however, the probability of scoring in any intermediate category is given by

$$P_{nij} = \frac{\exp(B_n-D_i-F_j)}{(1+\exp(B_n-D_i-F_j))} \quad \frac{\exp(B_n-D_i-F_{j+1})}{(1+\exp(B_n-D_i-F_{j+1}))} \qquad (7)$$

meaning that the probability of an observation in category j is the probability of succeeding on a dichotomous item associated with category j, less the probability of succeeding on one associated with category j-1. Thus, since Pnij>=0, then necessarily Fj+1>=Fj, so that the parameters for the ogives are monotonic with the category ordering.

This model has the desirable property that reversing the category numbering

maintains the scaling system, but it is not strictly a Rasch measurement model since it lacks sufficient statistics.

An advantage of this model is that redefining the rating scale by merging or splitting categories does not change the frame of reference. Consequently initial calibrations and measures can be estimated by bisecting the rating scale, forming a group of higher categories, and another group of lower categories, and then scoring all items as dichotomies based on which group the observed rating belongs to. More precise estimates can then be obtained by successively bisecting each of the groups, until the groups each comprise one category of the rating scale. The stability of the measures across bisections is an indication of the fit of the model.

This model is advantageous if the category boundaries are entirely arbitrary, so that instituting a category boundary at one position is just as good as another. Further, the category boundaries are independent, so that the presence or absence of one does not affect calibrations based on an adjacent boundary, apart from estimation considerations.

II. Modelling and communicating the intentional form of the scale

The rating scale models considered here, and others (e.g. Samejima 1972), have the drawback that the calibrations related to the categories may not be immediately comprehensible to the developer. To illustrate the problem and to provide a basis for some graphical solutions, Table 2 and Table 3 present the rating scale calibrations for two structurally similar rating scales analyzed using the Andrich model. The Figures in this section were excerpted from the output of the BIGSCALE (Wright et al. 1989) Rasch analysis computer program.

Consider a scale in which each category clearly represents a qualitatively higher level of the variable. The test developer is thinking in terms of Figure 1. The calibrations for such a scale are presented in numerical form in Table 2. The step calibrations in column 2 are in ascending numerical sequence and can be thought of as the transition points in Figure 2. In fact, the Andrich calibrations correspond to the points in Figure 5 at which the probability curve for each category intersects with the curve for the category below it, indicated by "+" signs.

Table 3 presents the calibrations for a less clearly defined scale. The Andrich step calibrations are no longer in sequence with the categories. The matching category probability curves are shown in Figure 6. The correspondence between the calibrations and the intersection points is still the same, being the points of intersection between the curve for one category and that for the category below it, marked by "+" signs. Since, however, each categories is not in turn the most probable, the curves do not form a procession of "hills". This means that the intersection points are disordered with respect to the category numbers. This disordering of intersection points is true for all rating scale models, but is reflected in different manners by the parameters.

Rather than consider the probability of any individual category, the cumulative probability curves, or "zone" curves (Masters 1980), corresponding to the

probability of observing that category or below could be drawn. These are shown in Figure 7 for the clearly-defined scale. Disordering of the Andrich calibrations is reflected by the close proximity of some of the cumulative curves in Figure 8. For the McCullagh model, these curves would always be parallel logistic ogives, the complement of Figure 4, and the points of intersection between the ogives and the .5 probability line would be the McCullagh category calibrations, a very clear pictorial representation of the scale. The Andrich and Glas models can also be depicted as plots of parallel logistic ogives, in which again the .5 probability line intersects each ogive at the category calibration point, but these plots are more tortuously related to the expected responses, and, through them, to the empirical scores.

An alternative approach to the scale is not in terms of the occurrence of any particular category, but in terms of what score the person is expected to make on the item. For the calibrations in Tables 2 and 3, these are shown by the ogives in Figures 9 and 10. For the models considered here, the score ogives must be monotonic ascending. They are calculated by summing the products of the category number and its probability for each point on the variable. Disordering of Andrich calibrations is reflected by marked changes in slope of the ogive. The sectors of the ogive, which, when rounded to the nearest integer, correspond to each of the possible expected scores on the item, are indicated by "|" bands. The "*" bands indicate the point at which an expected score exactly corresponds to a category number, the value of the expected score.

In many respects, Figure 1 is conceptualized by the developer in terms of observed score intervals, rather than category probabilities, so that the expected score bands in Figures 9 and 10 most closely correspond to the idealization in Figure 2. The numerical details of the integral expected score intervals are presented in the right-hand columns of Tables 2 and 3. These score interval calibrations can then be compared with the person measures and item calibrations to determine what category score the person is expected to achieve on the item. The score ogives for all the models have much the same appearance.

Incorporating the developer's intentions

The conventional Rasch-based analysis of Rating Scales is based on the premise that nothing is known, a priori, about the structure of the Rating Scale apart from the fact that numerically higher rating scale categories represent "more" of the latent variable. The general approach (e.g. MSCALE, 1986) is to collapse a scale into ascending ordinally counted categories and estimate the calibrations of the steps between the observed categories strictly on the basis of the observations in the data set at hand. For many applications this is sufficient to lead to useful calibration of the rating scale structure.

In examining and describing the models to this point, the relationship between the empirical data and the developer's intentions have been down-played. The empirical data, however, always depart to some extent from the developer's ideal. Consequently, parameter estimates obtained from the data only give an approximation to the rating scale structure. For instance, no observations of a particular category may occur at all in the data set under examination, though that category has good theoretical grounds to exist, and has been observed in

other data sets.

It is rarely the case in modelling rating scales that the developer is interested in describing a particular set of empirical data as precisely as possible. Rather, the empirical data which is being used to calibrate or validate the scale is known to be but one manifestation of the use to which the scale is to be put, and so long as the data does not markedly challenge the developer's intentions, it is those intentions which are superordinate, not the characteristics of the particular data set being analyzed. The next stage, therefore, is to constrain the step calibrations in accord with the developer's intentions, which are always an idealization of the scale. The fit of the data to the resulting model will indicate the degree to which the ideal is challenged by the actual.

The concept of modelling the developer's intentions by means of algebraic relationships between the category calibrations is well known. Rasch (1960/1980), Wright and Masters (1982), Masters and Wright (1984) present certain scale structures which could be regarded as rating scales, such as Poisson counts and counts of successful Bernoulli trials. The concept of choosing the most useful model is also well established. If it is not clear whether all the trials involve situations of equal difficulty then a decision must be made as to whether to fit the data to a Bernoulli model or a more general rating scale model.

Anchoring the scale calibrations

The most extreme distortion to a rating scale in an empirical data set occurs when a category is not observed. The missing category can be forced into the analysis and calibrated, merely by including in the analysis a dummy data record containing such an observation. This will lessen the distortion introduced into the frame of reference by the omission of the category, and so improve the overall quality of all the calibrations, but it is unlikely to lead to an accurate set of calibrations for the rating scale.

A more useful approach to distortion of the rating scale, for whatever reason, may be to pre-set or anchor the rating scale calibrations. If the rating scale is well understood, it is likely that a useful set of calibrations for the scale has already been obtained. These can be forced into the analysis, and the degree to which the data reflect the mandated structure can be determined by means of fit statistics and residuals. Some analysis software, e.g. FACETS (Linacre, 1988) and BIGSCALE, permit the pre-calibration, ("anchoring"), of category calibrations for both observed and unobserved categories.

More general structural concepts

Scale designers may intent to construct their scales in terms of "equal interval", "skewed", or "clustered" categories. Operationalizing this ideas mathematically, however, is a considerable challenge.

If the design of the rating scale was intended to meet some goal (e.g. the scale is to be "equal interval"), the analyst may wish to assert this in the scale calibrations, both to estimate such an "equal interval", and to force any conflict between the design intention and the observed data to manifest itself

in fit statistics. In this way, it can be determined whether the empirical data contradicts or supports the intended design of the scale.

Andrich (1978) includes in his discussion the case of rating scales with "equidistant thresholds". However, his thresholds are conceptualized in terms of the intersections of the probability curves, shown in Figure 5. These may not correspond to what the scale designer considers to be "equal interval" in terms of Figure 9. Nevertheless, once the external considerations have been reduced to a mathematical expression involving Rasch rating scale parameters, such as Andrich's equidistant threshold model, it is relatively straight-forward to construct estimation equations. Their form will be close to those given in Wright and Masters (1982). Suitable fit statistics can also be calculated to report how significantly the data diverge from the intended design model for the scale.

A sample design problem

An example is now presented of a number of ways in which a notionally "equal interval" scale could be parameterized for the Andrich model. The intention here is to indicate to the analyst the nature of the information needed from the designer in order to be able to put into explicit mathematical form the scale designer's conceptualization of the rating scale. A more complex design would yield an even greater number of possible mathematical realizations.

i) The ~~e~~~~q~~ual probability thresholds of adjacent categories are at equal intervals. (Andrich's equidistant threshold case).

Using the Rasch rating scale parameterization:

$$\log(P_{nij} / P_{nij-1}) = B_n - D_i - F_j \qquad (8)$$

in which $F_0 = 0$, and $\Sigma(F_j) = 0$, where $j=0,J$

Then an equal interval scale would be one in which, $(F_j - F_{j-1}) = C$, a constant across all j, except the extreme. Then

$$F_j = C((j-1) - (J-1)/2) \qquad (9)$$

Such a set of $(F_j, j=1,5)$ would be $-2, -1, 0, 1, 2$, producing Figure 11. The value, C, could be either pre-set or estimated from the data. If the thresholds according to the empirical data are very disordered, then the estimate of C could be negative, indicating that only the extreme categories are most probable to be observed.

ii) If the rating scale is intended to represent counts of successes on similar, exchangeable, tasks, then it can be represented by a Bernoulli trials model. The Bernoulli trial model for a 6 category scale yields a rating scale of probability structure shown in Figure 12, with parameter values, following Wright and Masters (1982 p.51), of $(F_j, j=1,5) = -1.61, -.69, 0, .69, 1.61$

iii) The maximum probability points of non-extreme categories are at equal intervals, i.e. the "hill tops" in Figure 13 are equally spaced. The condition is:

$$\frac{\delta}{\delta Xk} \left(\exp\left(j*X_k - \sum_{h=1}^{j} F_h\right)/(\text{standardizing factor})\right) = 0 \qquad (10)$$

when $X_k - X_{k-1} = C$, a constant across all k, except the extremes.

Such a set of $(F_j, j=1,5)$ is -2, -1.24, 0, 1.24, 2.

iv) The points on the variable where the expected score is equal to the category value are equally spaced, i.e. the "*" bands in Figure 14 are equally spaced. The condition is:

$$\sum_{j=1}^{J} \exp\left(j*X_k - \sum_{h=1}^{j} F_h\right)/(\text{standardizing factor}) = k \qquad (11)$$

when $X_k - X_{k-1} = C$, a constant across all k, except the extremes.

Such a set of $(F_j, j=1,5)$ is -2, -1.24, 0, 1.24, 2.

For the Andrich model, these points on the variable are also the points of maximum probability for the form modelled in (iii).

v) The equal intervals are intended to represent uniform spacing of the levels representing equal probabilities of being scored in or above a certain category. The condition is:

$$\sum_{j=0}^{k} \exp\left(j*X_k - \sum_{h=0}^{j} F_h\right)/(\text{standardizing factor}) = 0.5 \qquad (12)$$

when $X_k - X_{k-1} = C$, a constant across all k, except the extremes.

A set of parameters is $(F_j, j=1,5) = -2, -1.22, 0, 1.22, 2$, depicted in Figure 15.

For the Andrich model, these parameters are close to those for options (iii) and (iv). For this same constraint applied to the McCullagh model, the parameter values would be

$$(F_j, j=1,5) \sim -2, -1, 0, 1, 2.$$

vi) The half-point expected score thresholds are equally spaced, i.e. the "|" bands in Figure 16 are equally spaced. The condition is:

$$\sum_{j=1}^{J} \exp(j*X_k - \sum_{h=1}^{j} F_h)/(\text{standardizing factor}) = k + 0.5 \qquad (13)$$

when $X_k - X_{k-1} = C$, a constant across all k, except the extremes.

Such a set of $\{F_j, j=1,5\}$ is $-2, -1.77, 0, 1.77, 2$.

f) Conclusions and implications:

Analysis of rating scales has tended to ignore the intentions of the designer of the scale. Thus it is has not been possible to answer the question "Does the empirical data support or refute the hypothesis that the rating scale is functioning in accord with the intentions of its designer ?" The challenge to the analyst is to discern the designer's intentions and to convert them into the mathematical model, which can most usefully advance the understanding of the scale.

Bibliography

Andrich, D. (1978) A rating formulation for ordered response categories. Psychometrika 43(4): 561-573.

Likert, R. (1932) A technique for measurement of attitudes. Archives of Psychology 140: 1-55.

Linacre, J.M. (1988) FACETS computer program. Chicago: MESA Press.

Masters, G.N. (1980) A Rasch model for rating scales. Doctoral dissertation, University of Chicago.

Masters, G.N. & Wright B.D. (1984) The essential process in a family of measurement models. Psychometrika 49(4): 529-544.

Wright, B. D. & Masters, G.N. (1982) Rating Scale Analysis. Chicago: MESA Press.

Wright, B. D., Congdon, R., Schulz, M. (1986) MSCALE computer program. Chicago: MESA Press.

Glas, C.A.W. & Verhelst, N.D. (1989) Using the Rasch model for dichotomous data for analyzing polytomous responses. Paper presented at the Fifth International Objective Measurement Workshop, Berkeley, California.

McCullagh, P. (198^) Regression models for ordinal data. Journal of the royal Statistical Society B 42,2 pp. 109-142.

Samejima, F. (1972) A general model for free-response data. Psychometrika 37 (2) monograph supplement No. 18.

Rating (Score)	Category-Items 1	2	3	Length of item item "test"
0	0	N	N	1
1	1	0	N	2
2	1	1	0	3
3	1	1	1	3

Table 1. Scoring of a 4 category item according to the Glas model. "N" indicates that the category item is considered not to be administered.

CATEGORY LABEL	STEP CALIBR.	STEP ERROR	EXPECTED SCORE CALIBRATIONS STEP-.5	AT STEP	STEP+.5
0	NONE			EXTREME	-3.69
1	-3.50	.08	-3.69	-2.31	-1.19
2	-1.00	.04	-1.19	-.25	.69
3	.50	.03	.69	1.81	3.19
4	3.00	.20	3.19	EXTREME	

Table 2. Calibrations for an empirically clearly-defined rating scale fitted to the Andrich model. The step calibrations are in ascending sequence.

CATEGORY LABEL	STEP CALIBR.	STEP ERROR	EXPECTED SCORE CALIBRATIONS STEP-.5	AT STEP	STEP+.5
0	NONE			EXTREME	-2.23
1	-1.00	.07	-2.23	-1.51	-.84
2	-2.00	.06	-.84	.15	.98
3	3.00	.02	.98	1.53	2.13
4	.00	.10	2.13	EXTREME	

Table 3. Calibrations for an empirically ill-defined rating scale fitted to the Andrich model. The step calibrations are not in ascending sequence.

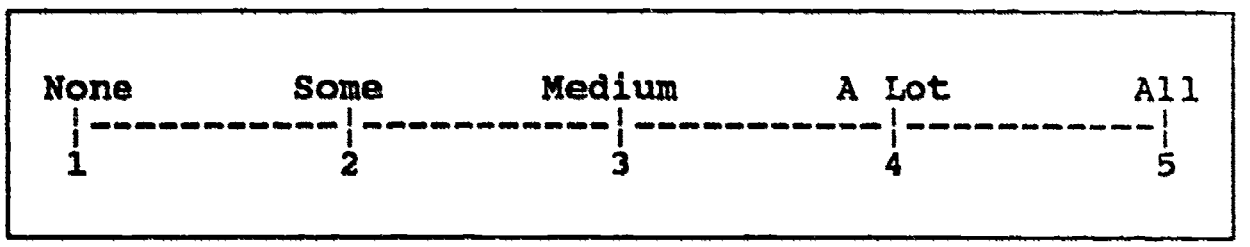

Figure 1. A scale developer's idealized conception of a rating scale.

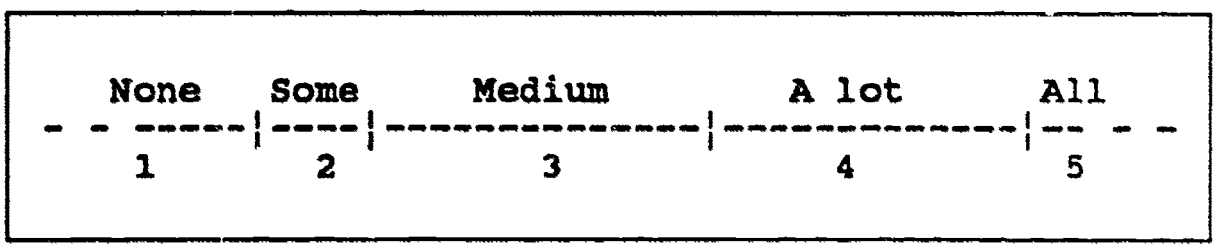

Figure 2. The rating scale expressed in terms of performance on the underlying variable.

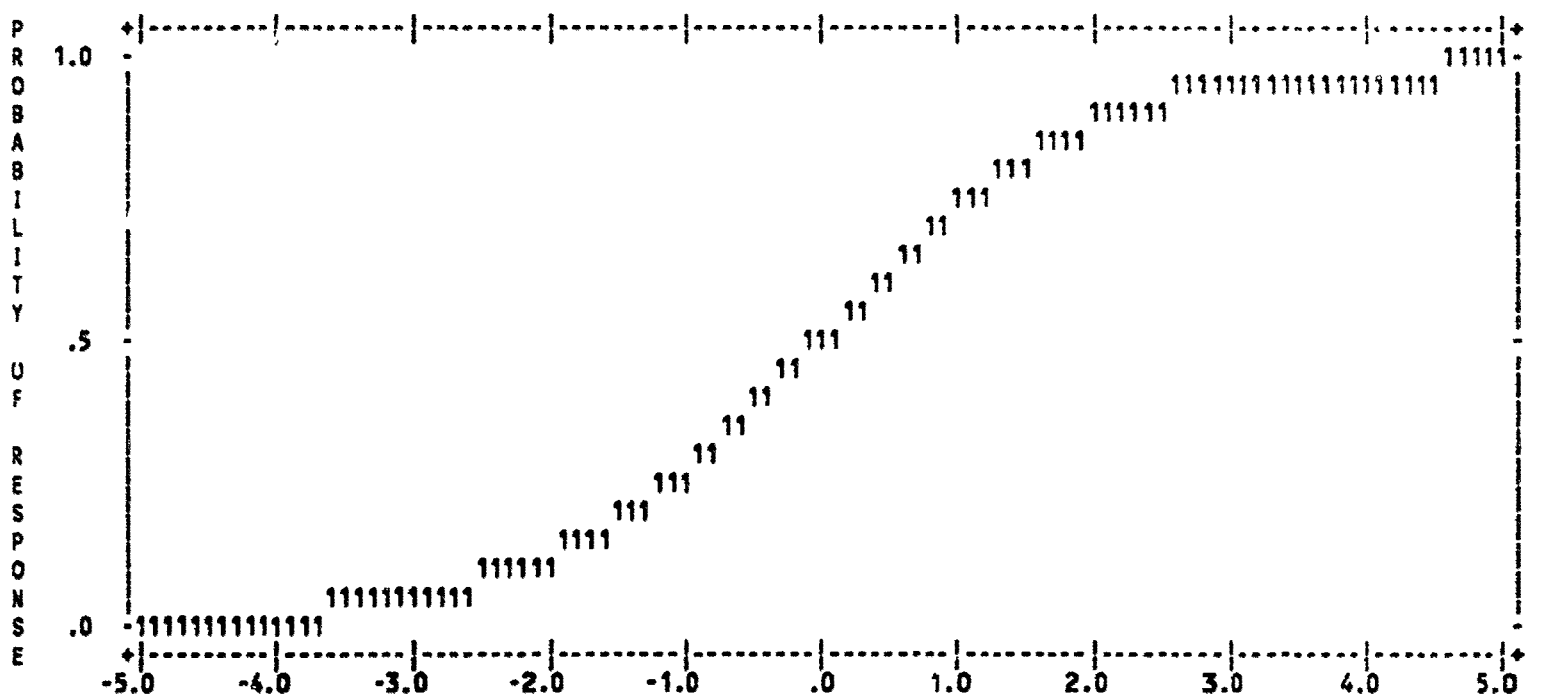

Figure 3. A simple ogive representing a dichotomy, expressing the relationship between a measure and an expected score.

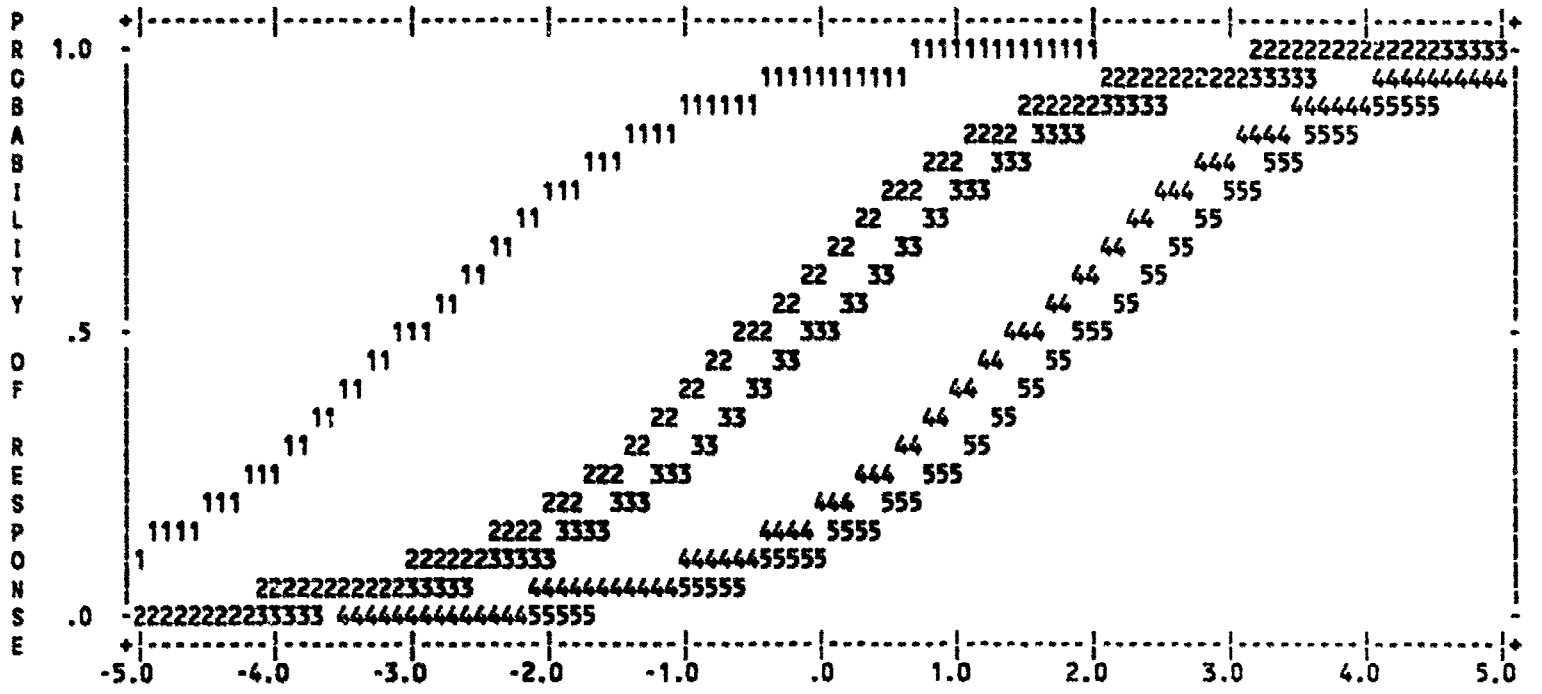

Figure 4. Parallel ogival curves. The means of obtaining measures from ratings.

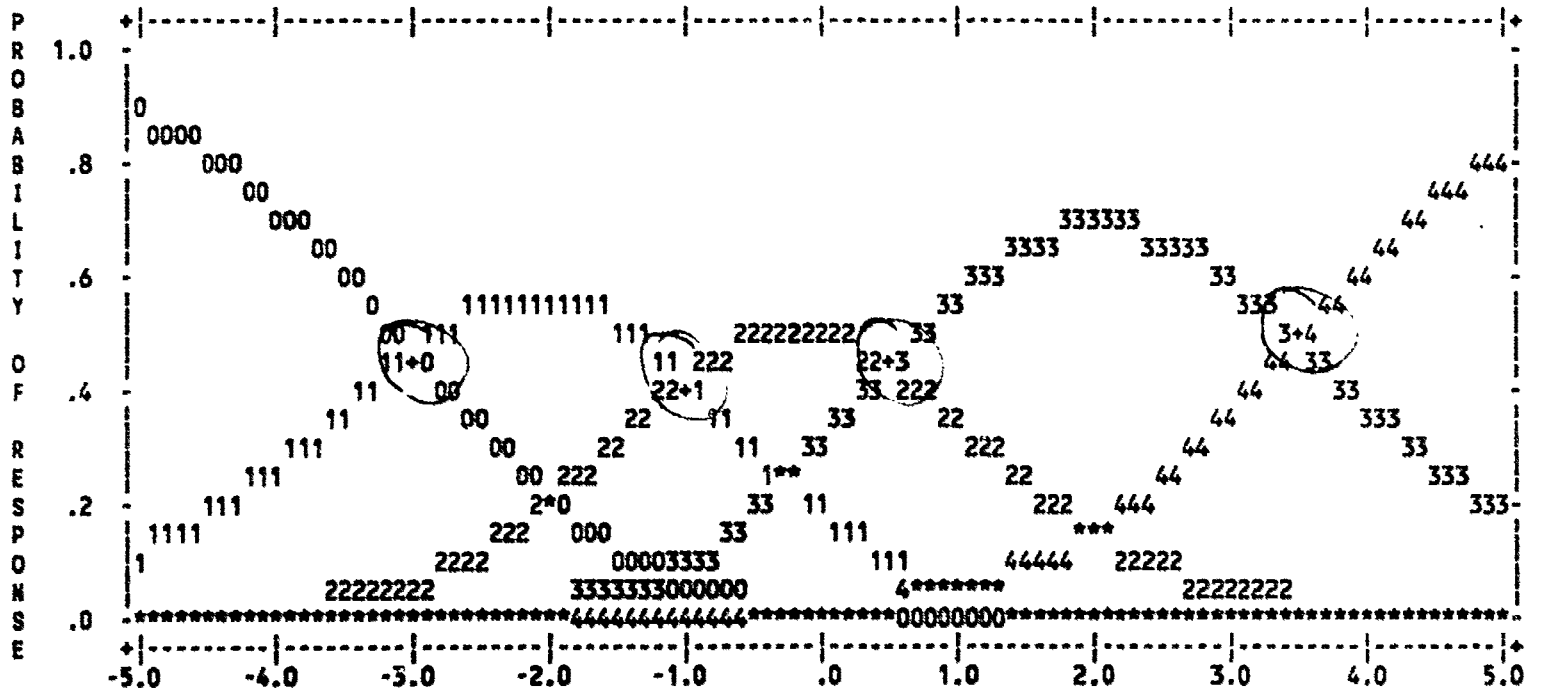

Figure 5. Probability of responding in each category of the clearly-defined Andrich scale for a person whose measure is indicated below the x-axis. The points of equal probability of adjacent categories are shown by "+".

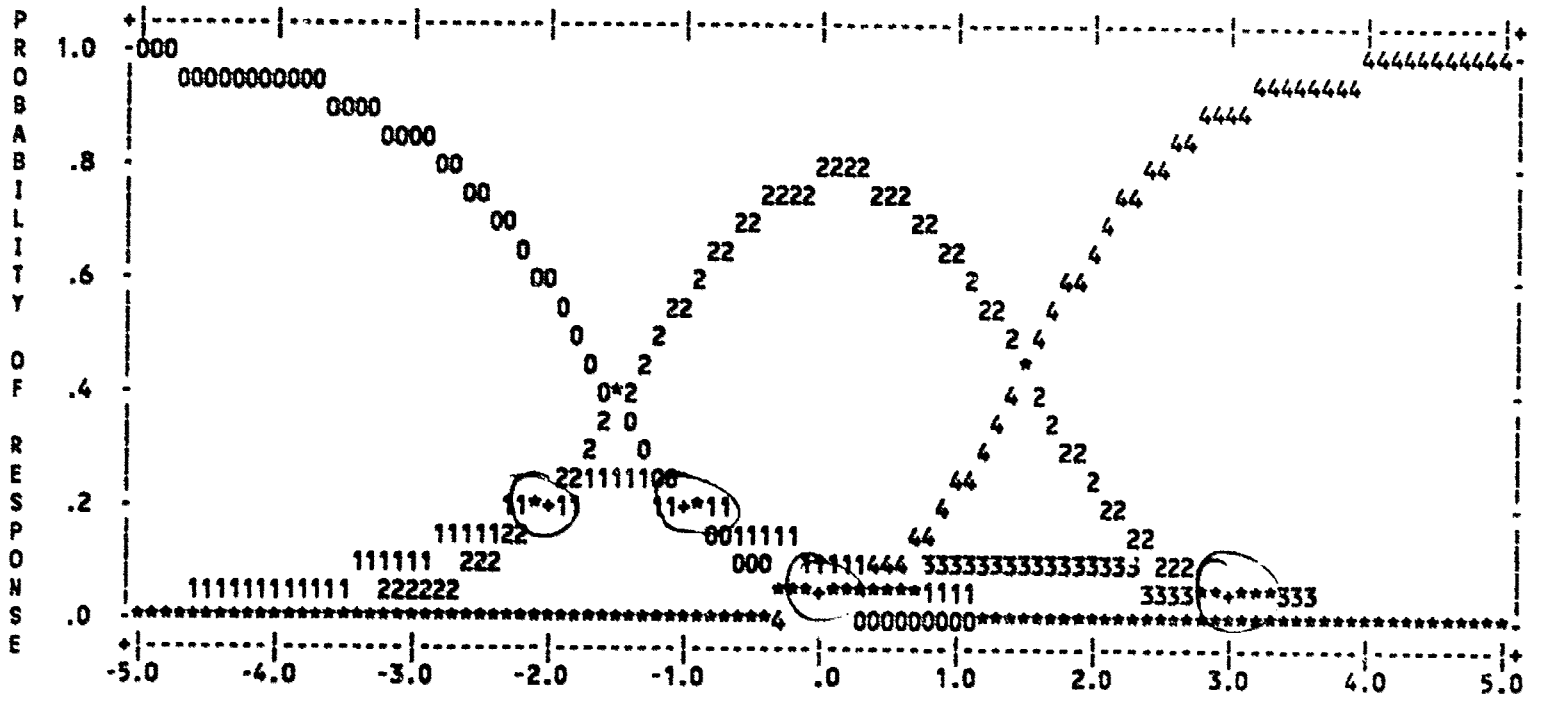

Figure 6. Probability of responding in each category of the ill-defined Andrich scale for a person whose measure is indicated below the x-axis. The points of equal probability of adjacent categories are shown by "+", and are disordered with respect to the categories.

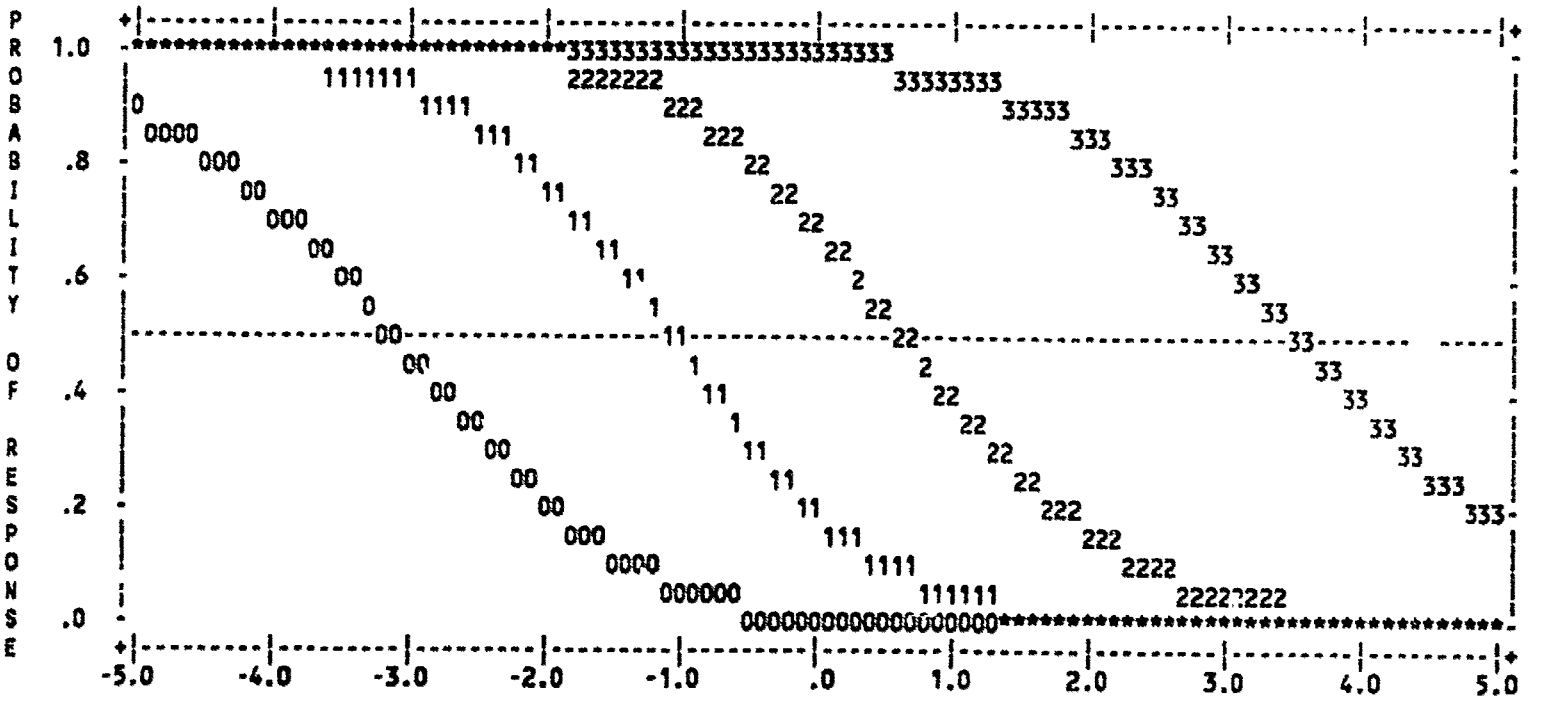

Figure 7. Probability of responding in a given category or below, for the clearly-defined scale, according to the Andrich model.

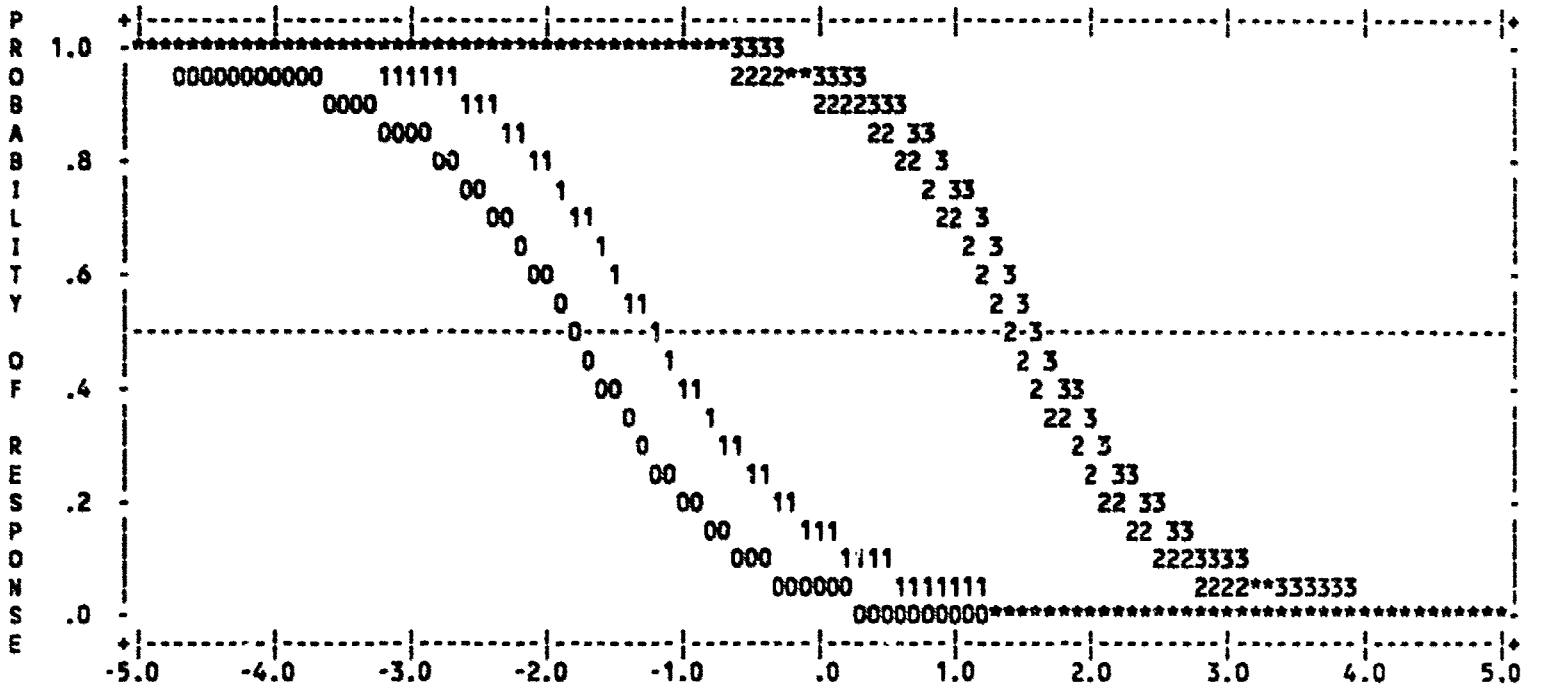

Figure 8. Probability of responding in a given category or below, for the ill-defined scale, according to the Andrich model.

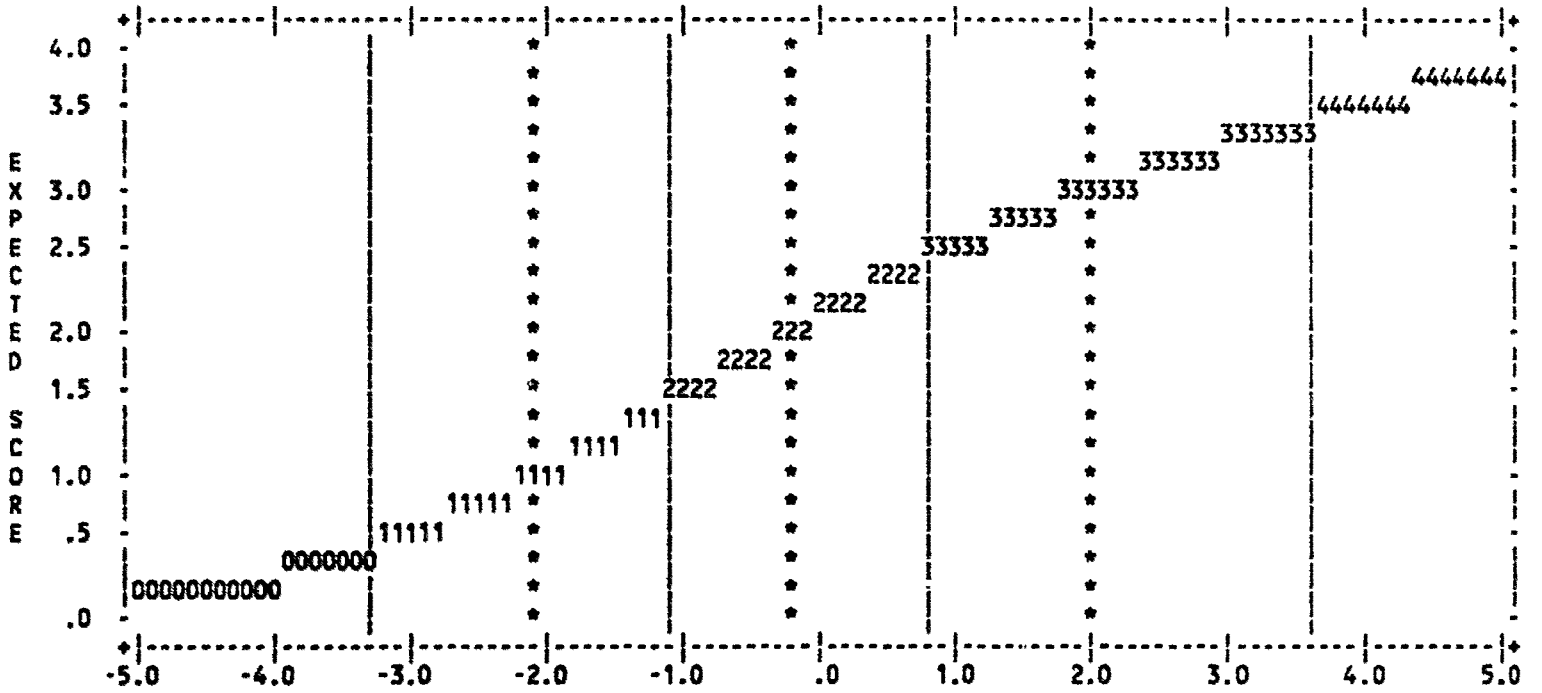

Figure 9. Expected score ogive for the clearly-defined scale.

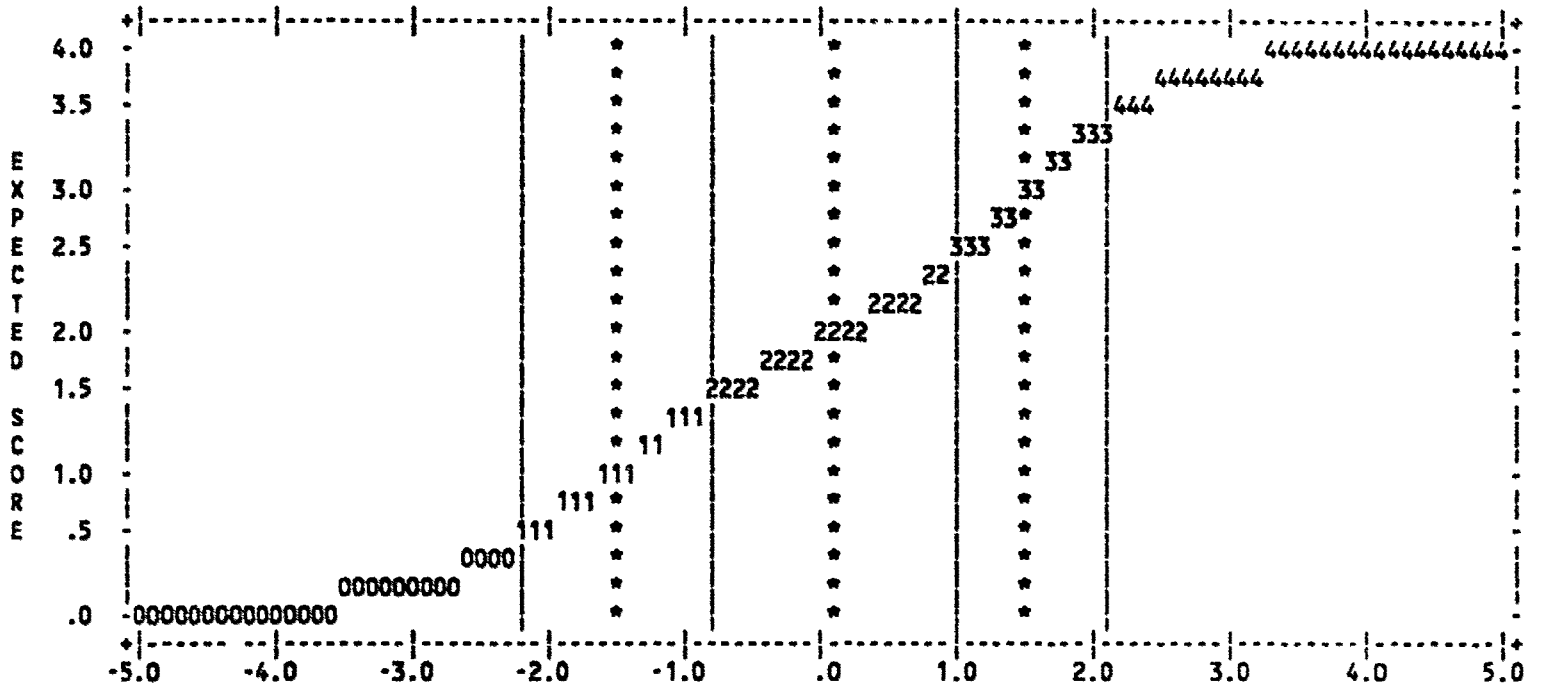

Figure 10. Expected score ogive for the ill-defined scale.

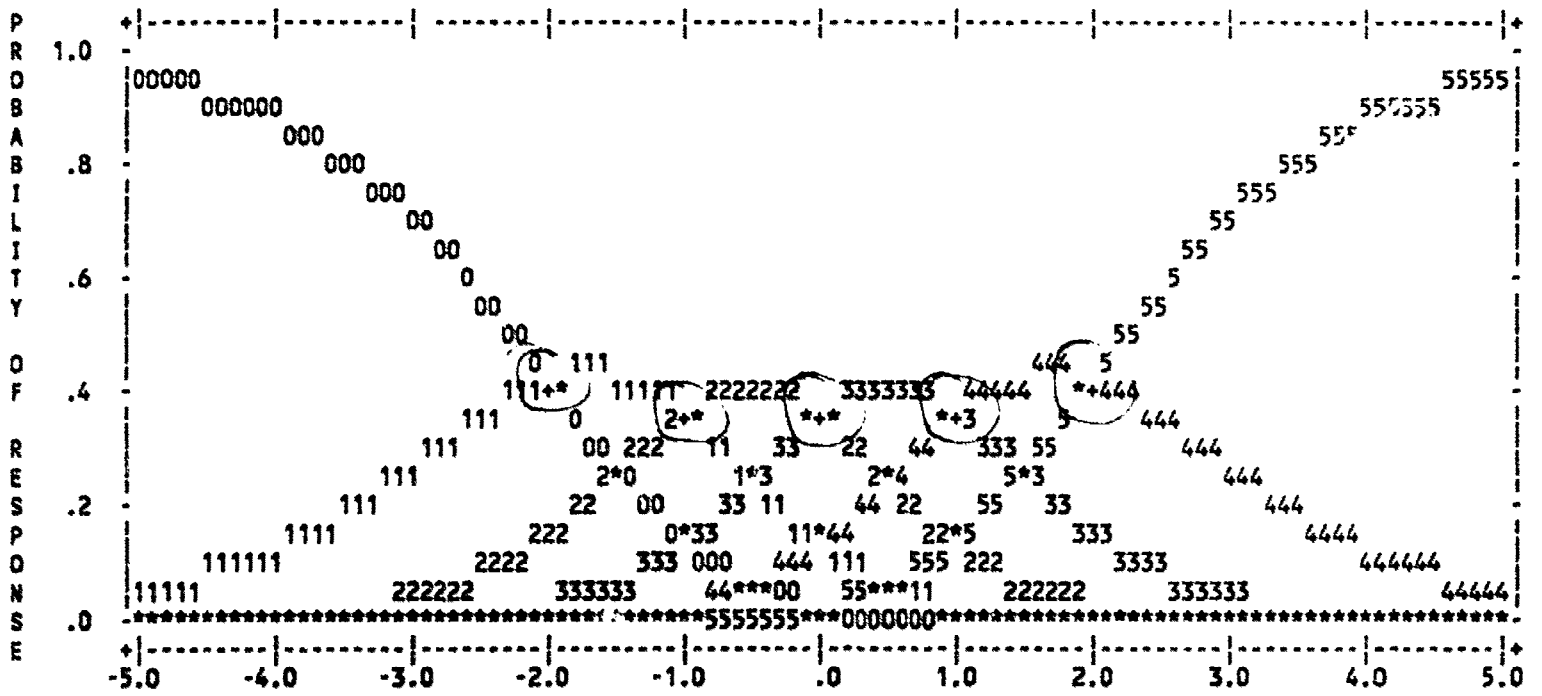

Figure 11. An "equal-interval" scale interpreted as equi-distant intersections of adjacent category probability curves. The intersections are indicated by "+".

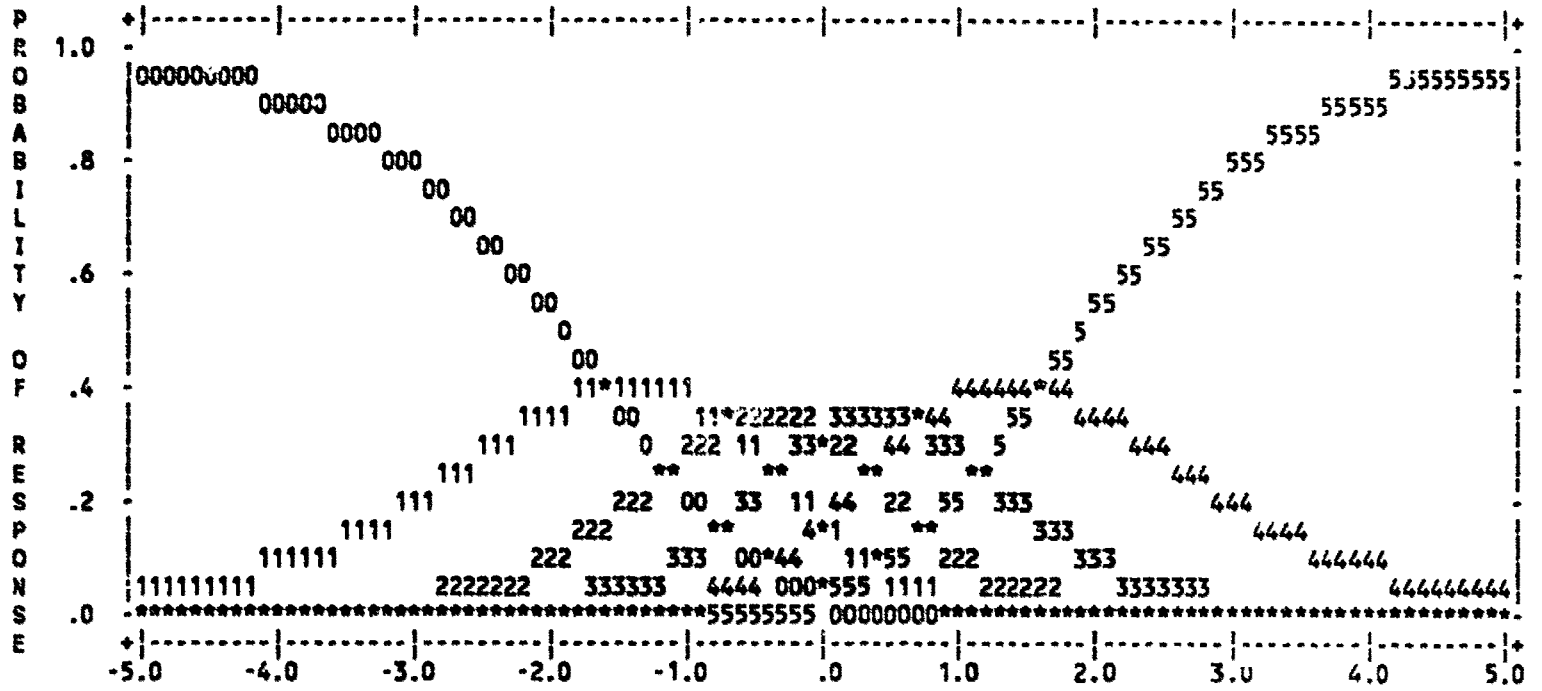

Figure 12. An "equal-interval" scale interpreted as counts of successes on Bernoulli trials.

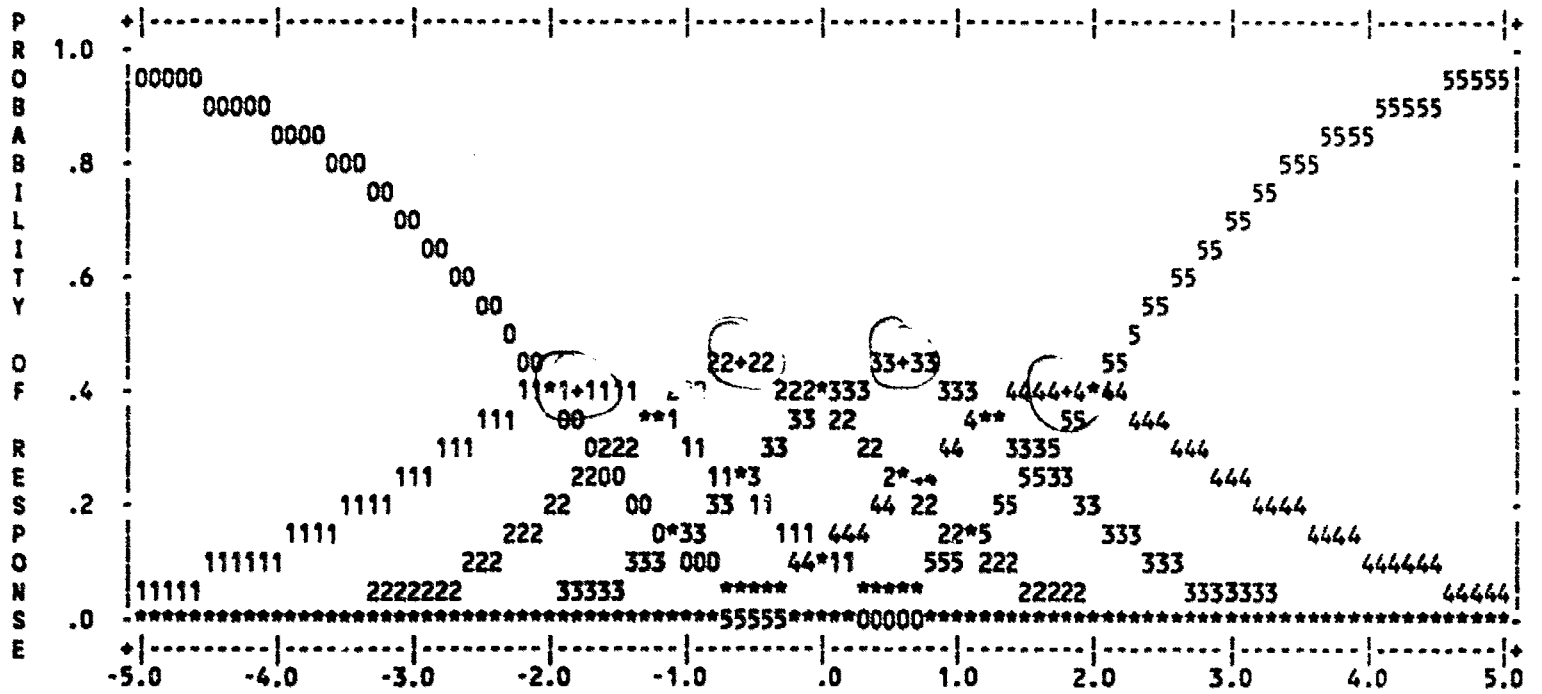

Figure 13. An "equal-interval" scale interpreted as equally spaced maxima of the intermediate category probability curves. The maxima are indicated by "+".

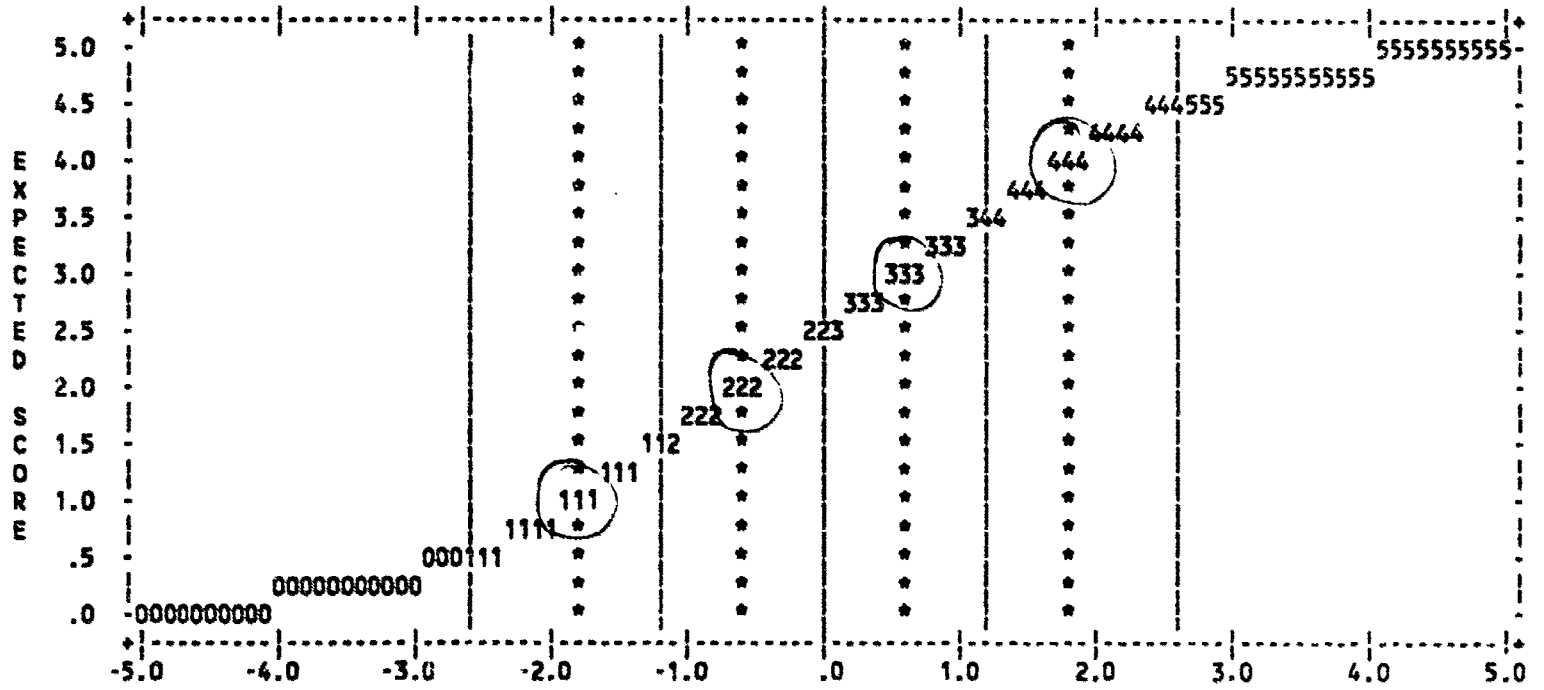

Figure 14. An "equal-interval" scale interpreted as equally spaced integral expected score values, which are indicated by the vertical "*" lines.

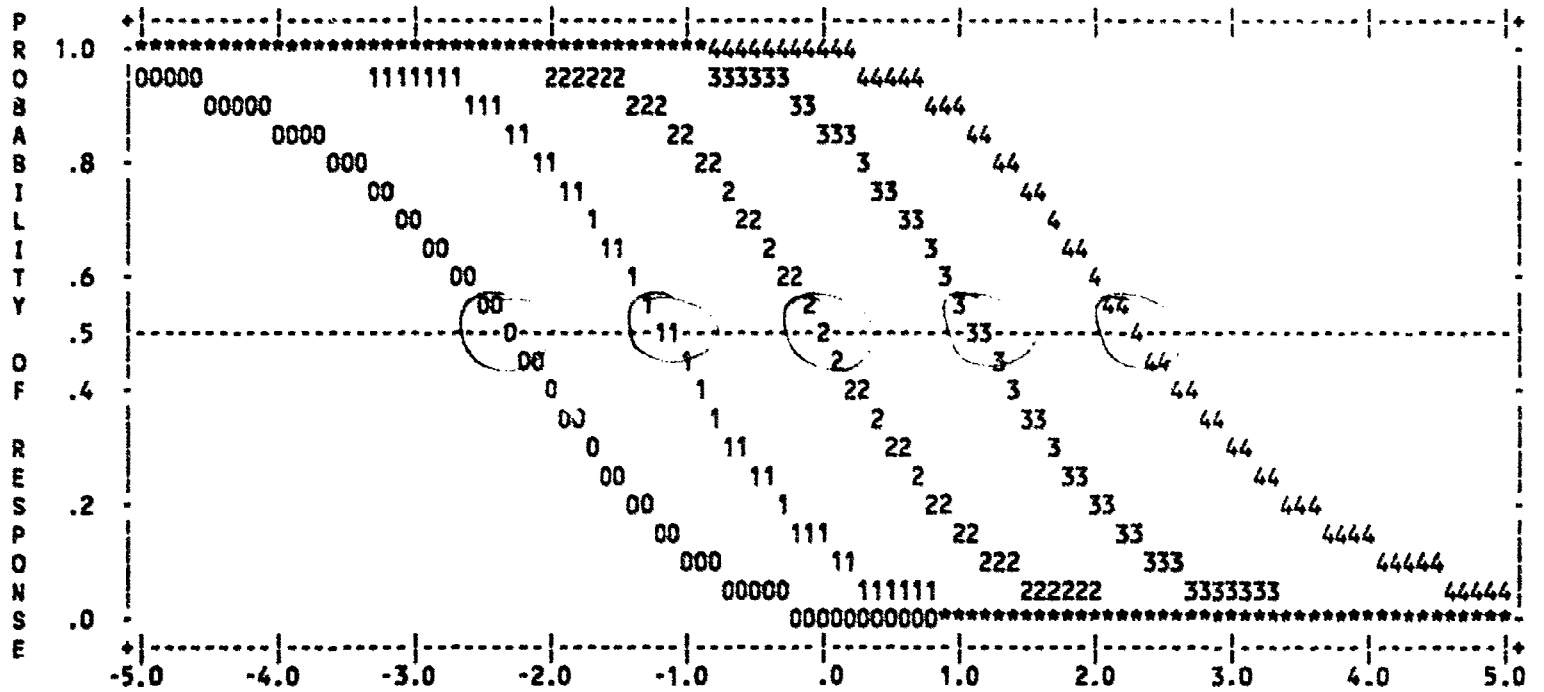

Figure 15. An "equal-interval" scale interpreted as equally spacing of the intersctions between the cumulative probability curves and the 0.5 probability line.

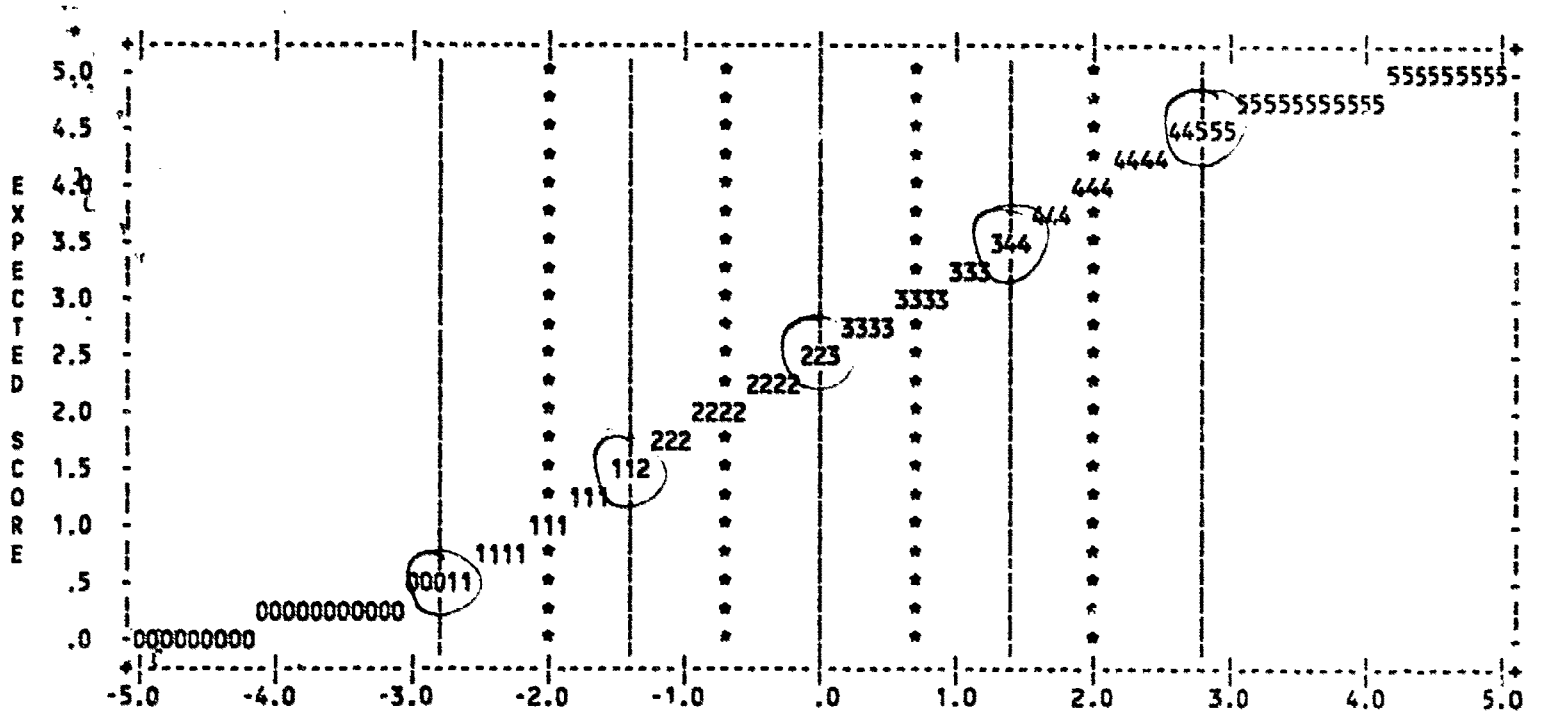

Figure 16. An "equal-interval" scale interpreted as equally spacing of the thresholds between adjacent rounded expected score intervals, indicated by "|" lines.